CONCLUSION

This work is primarily focused on the development of accurate and efficient model independent computational methods for genome annotation. Model independency of computational methods is a desirable feature particularly for newly discovered genomic sequence. After sequencing of the genomic data genome annotation is the first step that helps in the characterization that leads to the extraction of structural and functional features pertaining to the DNA sequences. Identification of coding sequences (CDS) is an important part of genome annotation. CDS determination is based on identification of exonic regions in eukaryotes. Exonic regions are important as they correspond to the protein coding regions. Sequences of amino acids in the exonic regions determine the type of protein formation. Signal Processing based computational methods for genome annotation are capable to provide model independent solutions. For this they depend on certain periodicities that characterize a particular genome region. Presence of period-3 component in the exonic region is exploited by Signal Processing based methods for CDS determination. However, with short exons or short exons disunited by short introns, weak period-3 component makes the problem more challenging. This problem has been addressed in this piece of research work by combining the results of multiple mappings by PCA. Identification of exon boundaries also known as splice sites is a problem associated with the CDS determination. The solution developed for the identification of short exons has been extended to mark the acceptor splice sites which correspond to the initiation of exonic regions. Apart from the application of computational methods for genome annotation they are also useful in evolutionary studies, DNA finger printing and identification of neuromuscular diseases and cancer. Presence of tandem repeats and mutations, insertions and deletions in them are utilized for these applications. Model independent computational methods analyze the tandem repeats with the help of frequency present in these regions. Time-frequency tools are highly useful as they can search the temporal locations corresponding to the frequency components. Amongst the three types of tandem repeats namely, satellites, minisatellites, and microsatellites the last one, also known as short tandem repeats, is most important. Microsatellites are characterized by the presence of period 2 to 6 periodicities. In this work optimized S-transform has been used for their identification. An application of the developed algorithm in the identification of repeat...
expansion diseases has also been demonstrated. Further improvisation in this algorithm helped us to identify minisatellites along with microsatellites.

The four algorithms developed in this work have a potential for further improvisation and expansion. Also, this work opens up new vistas to pursue research in the following areas-

(a) Identification CpG islands
(b) Identification of solenoid repeats
(c) Identification of transcription starts sites
(d) Identification of UTRs
(e) Cancer Genome Research